

PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (<http://bmjopen.bmj.com/site/about/resources/checklist.pdf>) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Predicting the Hand, Foot, and Mouth Disease Incidence Using Search Engine Query Data and Climate Variables: An Ecological Study in Guangdong, China
AUTHORS	Du, Zhicheng; Xu, Lin; Zhang, Wangjian; Zhang, Dingmei; Yu, Shicheng; Hao, Yuantao

VERSION 1 - REVIEW

REVIEWER	Jaranit Kaewkungwal Faculty of Tropical Medicine, Mahidol University, Thailand
REVIEW RETURNED	01-Mar-2017

GENERAL COMMENTS	<p>The study was written clearly in terms of its objectives and statistical models used. The incomplete parts are in the data preparation and methodology in data management for analysis. Such information should be described to support the validity and reliability of study results and discussion. Major comments are as follows:</p> <ol style="list-style-type: none">1. What was the quality of the data from the three sources? Were there any missing data? If so, how the researcher handled such problem.2. Were there any null values (no cases), both surveillance and BDI data, in any particular week? If so, how the researchers manage it?3. It is not clearly written how temperature per week was calculated for the whole province.4. It would be more informative for readers if the researchers also described what and how BDI units were counted; i.e., what/how were the keywords or searches counted?5. There appeared to be a major outbreak in 2014 (as shown in Figure 1). How did the researcher handle such event?6. The researcher did test for goodness of fit and precision of prediction as mentioned in the text. It was not clearly written, however, what or how the data were used to develop the model and to test the models. As shown in Figure 4 regarding forecast accuracy, is it the data from 200 weeks used to develop model and 8 weeks to test the model? But the researcher mentioned in Figure 4 that it is the forecasting of one quarter of HFMD incidence.7. It is somewhat acceptable to consider temperature as a climate variable. Even though it has been observed and noted that HFMD incidences are high in summer and autumn in several countries and almost throughout the year in tropical countries, but please give rationale(s) why temperature would affect the HFMD incidence. The lag of 1-week lag came from analytic model but is there any biological or clinical or hygienic reason for it? This may help to make a stronger discussion.8. The study focused mainly on methodology/statistical modelling. However, based on the results of the study, could the researcher discuss on how BDI would be useful information for public health
-------------------------	--

	use in monitoring/prevention/control of HFMD as claimed?
--	--

REVIEWER	Cordia Chu Griffith University, Australia
REVIEW RETURNED	22-Mar-2017

GENERAL COMMENTS	<p>This paper aims to examine the utility of temperature and search engine query data in predicting risk of Hand-Food-Mouth Disease (HFMD) which can enhance the HFMD surveillance system in China. Overall, the study has proposed an innovative approach using existing data to predict HFMD risk, and the findings may make important contribution to HFMD prevention and control. However, there are some concerns that the authors should consider in order to improve the quality of the paper.</p> <p>My comments are below:</p> <p>Major comments:</p> <ol style="list-style-type: none"> 1. It is not clear why the authors used "Distributed Lag Non-Linear Mode" to examine the relationship between predictors and HFMD incidence and then used a different model (ARIMAX) to develop the prediction model. Since these two models have different statistical characteristics and prediction ability, the authors need to spell out why they do it this way? 2. Likewise, why did the authors use AIC to validate the model with BDI but used MAPE to validate model with BDI + Temperature. Thus, the validation methods used were not consistent. The authors should address this point with caution 3. The time-span (e.g. 3 weeks, 6 weeks, etc.) that prediction model can predict HFMD risk in advance need to be examined and reported clearly 4. How the Baichi Index contributes to the prediction of HFMD risk should be clearly described 5. The authors should have more detailed discussions about how the proposed prediction model can play a role in the existing surveillance system, and then highlight the significance of this work. <p>Minor comments:</p> <ol style="list-style-type: none"> 1. Introduction, Page 3 Lines 25-28: BAIDU needs to be explained in full rather than referring readers to the website 2. Method, Page 5 Line 1: the same comment as 1 3. Method should have Data Analysis as a sub-heading before coming to a lower-level subheading: Descriptive analysis, DLNM, ARIMAX. Data Analysis should have some lines to outline the methods applied in the study 4. Results, Page 6 Line 14: "Descriptive analysis" should be replaced by "Descriptive statistics" 5. Results, Page 7 Line 6-7: what is the threshold temperature in non-linear relationship between temperatures and HFMD incidence?
-------------------------	--

REVIEWER	Byung Chul Chun Korea University College of Medicine, Seoul, Korea
REVIEW RETURNED	04-Apr-2017

GENERAL COMMENTS	<ol style="list-style-type: none"> 1. It is necessary to add a description of Baidu index (BDI) for HFMD to the research methods. I understand BDI is a Chinese counterpart of Google Trend. Please explain how the BDI for HFMD was created for this research; what query(or search) terms did you use to get the
-------------------------	---

	<p>BDI for HFMD in Guangdong?</p> <p>2. Fig 2 shows the BDI tends to increase over time. What makes this increase of BDI? Is there internet penetration changes between 2011 and 2014 in Guangdong? What are the internet penetration rate in the study period by year? Add an explanation of BDI changes over study period and its possible effects on your results in discussion session.</p>
--	---

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

The study was written clearly in terms of its objectives and statistical models used. The incomplete parts are in the data preparation and methodology in data management for analysis. Such information should be described to support the validity and reliability of study results and discussion. Major comments are as follows:

1. What was the quality of the data from the three sources? Were there any missing data? If so, how the researcher handled such problem.

Response: Thanks for the positive comments. There was no missing data involved in our study. We have revised the section of ‘data collection’, as follows, to clarify the quality of our data: (Methods section, page 4)

“Weekly Case-based HFMD surveillance data from 2011 to 2014 were obtained from the National Center for Public Health Surveillance and Information Services, China Center for Disease Control and Prevention (China CDC). HFMD disease was made statutorily notifiable since the large outbreak in May 2008 and the enhanced national surveillance system has been described and validated in detail elsewhere [15]. Weekly incidence of HFMD was calculated using data from the Annual Statistical Report of Guangdong (<http://www.gdstats.gov.cn/tjsj/gdtjnj/>). Daily meteorological data were obtained for the same period from the China Meteorological Data Sharing Service System (<http://cdc.nmic.cn/home.do>) which is the oldest and authorized meteorological department in China. Data of daily temperature from six monitoring stations in Guangdong were aggregated into weekly average data for the whole province. Daily search engine query data were obtained for the same period from the Baidu Inc. (<https://index.baidu.com/>), which is the one of the largest Internet companies in the world [13]. The Baidu index has made great value for market monitoring, media selection, advertising and effect evaluation [13]. We searched Baidu using the keyword of ‘hand, foot, and mouth disease’ and counted the search frequency recorded by the Baidu Inc. Daily search frequencies were also aggregated into weekly average data. All information above was collected from Jan 2011 to Dec 2014, a total of 208 weeks, and was included in the data analysis. No missing data were observed.”

2. Were there any null values (no cases), both surveillance and BDI data, in any particular week? If so, how the researchers manage it?

Response: Thanks for the comments. There was no null value involved in our study. To avoid further confusion caused, we have added the minimum and the maximum value of each variable in the Table 1, as follows: (Results section, table 1, page 16)

Variables	Minimum	Median	Maximum
Incidence (1/10,000)	0.018	0.482	2.428
Temperature (°C)	8.795	23.07	29.893
BDI (unit)	516	3296	32596

3. It is not clearly written how temperature per week was calculated for the whole province.

Response: Our apology for not being clear. We have added the description of the calculation of temperature in the Methods section, as follows:

“Data of daily temperature from six monitoring stations in Guangdong were aggregated into weekly

average data for the whole province.”

4. It would be more informative for readers if the researchers also described what and how BDI units were counted; i.e., what/how were the keywords or searches counted?

Response: Thanks for the comments. We have added the following description in the Methods section:

“We searched Baidu using the keyword of ‘hand, foot, and mouth disease’ and counted the search frequency recorded by the Baidu Inc. Daily search frequencies were also aggregated into weekly average data.”

5. There appeared to be a major outbreak in 2014 (as shown in Figure 1). How did the researcher handle such event?

Response: Thanks for the question. To handle the major outbreak, we have conducted two differencing steps before fitting the model (please refer to the Results section, page 8). In addition, we have used the Ljung-Box test to test whether the residuals of the model were independent of the white noise (please refer to the Methods section, page 6).

6. The researcher did test for goodness of fit and precision of prediction as mentioned in the text. It was not clearly written, however, what or how the data were used to develop the model and to test the models. As shown in Figure 4 regarding forecast accuracy, is it the data from 200 weeks used to develop model and 8 weeks to test the model? But the researcher mentioned in Figure 4 that it is the forecasting of one quarter of HFMD incidence.

Response: Our apology for not being clear. We have added the following information in Methods to clarify:

“In our study, data for a total of 208 weeks were available and used to both develop the ARIMAX models and test the goodness of fit. To estimate the forecast accuracy of the models, we divided the data into two data sets, one for training (195 weeks) and the other for testing (13 weeks, about one quarter of a year).”

7. It is somewhat acceptable to consider temperature as a climate variable. Even though it has been observed and noted that HFMD incidences are high in summer and autumn in several countries and almost throughout the year in tropical countries, but please give rationale(s) why temperature would affect the HFMD incidence. The lag of 1-week lag came from analytic model but is there any biological or clinical or hygienic reason for it? This may help to make a stronger discussion.

Response: Thanks for the comments. We have added the information as suggested. (Discussion section, page 9)

“Our finding of the lag of 1-week was well consistent with the incubation period of enteroviruses and the potential delay in parental awareness [23], which was also reported in a previous study in Hong Kong [24].”

“Previous laboratory studies showed that under 35°C, higher temperature was associated with a higher survival rate of enteroviruses [26, 27]. Furthermore, higher temperature might also lead to more frequent outdoor activities, which would subsequently lead to a higher risk of HFMD.”

8. The study focused mainly on methodology/statistical modelling. However, based on the results of the study, could the researcher discuss on how BDI would be useful information for public health use in monitoring/prevention/control of HFMD as claimed?

Response: Thanks for the suggestions. We have added the following information in the Discussion: “Guangdong is one of the most developed provinces of China with a high Internet penetration (72.4%) [17]. As Baidu Inc. is the largest Chinese search engine, its search queries could be a good representative of the needs of people’s lives, especially in regions with high Internet penetration. Moreover, as search queries can be processed quickly, applying the ARIMAX model with real time BDI may provide opportunity for monitoring and early detection of HFMD, and become an important

front line of defense against future HFMD epidemics in Guangdong, and perhaps eventually in national settings.”

Reviewer: 2

This paper aims to examine the utility of temperature and search engine query data in predicting risk of Hand-Food-Mouth Disease (HFMD) which can enhance the HFMD surveillance system in China. Overall, the study has proposed an innovative approach using existing data to predict HFMD risk, and the findings may make important contribution to HFMD prevention and control. However, there are some concerns that the authors should consider in order to improve the quality of the paper.

My comments are below:

Response: Thanks for the positive comments.

Major comments:

1. It is not clear why the authors used “Distributed Lag Non-Linear Model” to examine the relationship between predictors and HFMD incidence and then used a different model (ARIMAX) to develop the prediction model. Since these two models have different statistical characteristics and prediction ability, the authors need to spell out why they do it this way?

Response: Our apology for not being clear. We have added the following information in the Methods section to explain why we used the distributed lag non-linear model before the ARIMAX model:

“We used the distributed lag non-linear model with quasi-Poisson distribution to collect the relative risk (RR) between predictors and HFMD incidence and to provide evidence for setting the lag of the predictors in the ARIMAX model.”

2. Likewise, why did the authors use AIC to validate the model with BDI but used MAPE to validate model with BDI + Temperature. Thus, the validation methods used were not consistent. The authors should address this point with caution

Response: Thanks for pointing out the unclearness. To clarify, we have added the following information in the Methods section:

“The fitness of the models was assessed by Akaike Information Criterion (AIC) to indicate the fitness when developing models, with a lower AIC value indicting a better goodness of fit. While the mean absolute percentage error (MAPE) was used to assess the prediction accuracy during forecasting, with a lower value indicting a more accurate prediction.”

3. The time-span (e.g. 3 weeks, 6 weeks, etc.) that prediction model can predict HFMD risk in advance need to be examined and reported clearly

Response: Our apology for not being clear. In our study, we divided the data set into two data sets, one for training (195 weeks) and the other for testing (13 weeks, about one quarter of a year) to estimate the forecast accuracy of the models. The MAPE values exceeded 100 for all four ARIMAX models including models 1) based on surveillance data only, and models with addition of 2) temperature, 3) BDI and 4) both temperature and BDI in the thirteenth week (MAPE values ranged from 102 to 123; Supplement table 1), indicating a poor forecasting performance after 13 weeks. Thus, the prediction model can predict HFMD risk up to 13 weeks in advance. We have added the following information to point out the time-span that the prediction model can predict in the Methods: “To estimate the forecast accuracy of the models, we divided the data into two data sets including the training data set (195 weeks) and the testing data set (13 weeks, about one quarter of a year). The MAPE values exceeded 100 for all four ARIMAX models including models 1) based on surveillance data only, and models with addition of 2) temperature, 3) BDI and 4) both temperature and BDI in the thirteenth week (MAPE values ranged from 102 to 123; table not shown), indicating a poor forecasting performance after 13 weeks. Thus the model predicts the HFMD risk up to 13 weeks in advance.”

Supplement table 1. Forecast accuracy of ARIMAX models.

Weeks	Model1	Model2	Model3	Model4
1	14.330	14.356	12.132	12.418
2	23.220	22.932	22.467	21.617
3	39.089	38.630	38.035	36.513
4	44.132	43.607	42.928	41.140
5	47.527	46.962	46.211	44.187
6	55.595	54.908	52.894	50.410
7	63.774	62.911	59.700	56.626
8	70.650	69.635	65.493	61.893
9	72.461	71.462	67.165	63.580
10	75.201	74.146	69.394	65.633
11	86.232	84.880	78.065	73.422
12	102.373	100.625	91.884	86.107
13	122.544	120.270	109.026	101.745

Model1: empty; Model2: temperature; Model3: BDI; Model4: temperature+BDI.

4. How the Baidu Index contributes to the prediction of HFMD risk should be clearly described.

Response: Thanks for the comment. We have added the following information to point out how the BDI contributes to the prediction in the Discussion section:

“Guangdong is one of the most developed provinces of China with a high Internet penetration (72.4%) [17]. As Baidu Inc. is the largest Chinese search engine, its search queries could be a good representative of the needs of people's lives, especially in regions with high Internet penetration. Moreover, as search queries can be processed quickly, applying the ARIMAX model with real time BDI may provide opportunity for monitoring and early detection of HFMD, and become an important front line of defense against future HFMD epidemics in Guangdong, and perhaps eventually in national settings.”

5. The authors should have more detailed discussions about how the proposed prediction model can play a role in the existing surveillance system, and then highlight the significance of this work.

Response: Thanks for the suggestions. We agree and have added the following information to point out how our model contributes to the existing surveillance system. (Discussion section, page 10)

“The current surveillance system would benefit from taking into account both BDI and temperature within the ARIMAX model, becoming more timely and efficient. Due to the lack of manpower and material resources, the surveillance system of China is relatively inefficient at the current stage [30]. Cases were reported through a stepwise hierarchical reporting system, i.e., from the lowest to the highest hierarchy at a sequence of town, county, city, province, and the national CDC. The proposed model incorporating the timely Internet search engine queries in our study was shown to improve the forecasting ability of the surveillance system substantially. Similar study by Ginsberg J et al. introduced...”

Minor comments:

1. Introduction, Page 3 Lines 25-28: BAIDU needs to be explained in full rather than referring readers to the website

Response: The term ‘Baidu’ is the full name of the Internet company. We have added more detailed information about the Baidu in the Introduction, as follows:

“A similar search index of Baidu Inc. (Baidu index, BDI: <http://index.baidu.com>), which is the world's largest Chinese search engine and the biggest Chinese website...”

2. Method, Page 5 Line 1: the same comment as 1

Response: Thanks for the comment. The term ‘Baidu’ has been explained in its first appearance in the paper (please refer to the response to the Minor comments 1 above).

3. Method should have Data Analysis as a sub-heading before coming to a lower-level subheading: Descriptive analysis, DLNM, ARIMAX. Data Analysis should have some lines to outline the methods applied in the study

Response: Thanks for the comments. We have revised the Methods by providing some sub-headings accordingly. The revised section of Data Analysis is as follows:

“Data Analysis

First, we used the descriptive analysis to present the minimum, median, maximum, mean and standard deviation (SD) of the HFMD incidence, as well as the time series of weekly incidence and weekly average temperature and BDI. Then we used the distributed lag nonlinear model to examine the association between predictors and HFMD incidence. Finally, we fitted the ARIMAX model to forecast the HFMD incidence. All data analysis was conducted in R (version 3.3.2), using packages including ‘base’, ‘psych’, ‘lattice’, ‘TSA’, and ‘dlnm’.”

4. Results, Page 6 Line 14: “Descriptive analysis” should be replaced by “Descriptive statistics”

Response: Thanks. We have revised it in the text accordingly.

5. Results, Page 7 Line 6-7: what is the threshold temperature in non-linear relationship between temperatures and HFMD incidence?

Response: Thanks for the question. We have added the related information in the Results section, as follows:

“The 1-week lag-specific lines of temperature-incidence association and the reference line (RR=1.0) overlapped at around 24 degrees Celsius.”

Reviewer 3:

Please leave your comments for the authors below

1. It is necessary to add a description of Baidu index (BDI) for HFMD to the research methods. I understand BDI is a Chinese counterpart of Google Trend. Please explain how the BDI for HFMD was created for this research; what query (or search) terms did you use to get the BDI for HFMD in Guangdong?

Response: Thanks for the positive and very helpful comments. We have added the following information to the Methods section. (page 5)

“We searched Baidu using the keyword of ‘hand, foot, and mouth disease’ and counted the search frequency recorded by the Baidu Inc. Daily search frequencies were also aggregated into weekly average data.”

2. Fig 2 shows the BDI tends to increase over time. What makes this increase of BDI? Is there internet penetration changes between 2011 and 2014 in Guangdong? What are the internet penetration rate in the study period by year? Add an explanation of BDI changes over study period and its possible effects on your results in discussion.

Response: Thanks for pointing out this issue. The BDI in our study indicates the search frequencies of ‘hand, foot, and mouth’. According to the China Internet network information center (CNNIC), the Internet penetration was 60.4%, 63.1%, 66.0% and 68.5%, respectively, for the year from 2011 to 2014. Hence, the increase of BDI could be partly due to the increase of the Internet penetration. However, we could not distinguish the increase in BDI due to the true increase in HFMD cases from the increase due to internet penetration. We have added this limitation in the Discussion section as follows:

“Finally, the increase of BDI may be partly due to the increase in Internet penetration in China during the same period. According to the China Internet network information center (CNNIC), the Internet penetration from 2011 to 2014 was 60.4%, 63.1%, 66.0% and 68.5%, respectively, each year [32]. However, we could not distinguish the increase in BDI due to the true increase in HFMD cases from

the increase due to internet penetration. Thus the overestimation of BDI, if any, may lead an overestimation of the association between BDI and HFMD incidence.”

VERSION 2 – REVIEW

REVIEWER	Jaranit Kaewkungwal Faculty of Tropical Medicine, Mahidol University, Thailand
REVIEW RETURNED	23-May-2017

GENERAL COMMENTS	The authors had revised the manuscript accordingly. It is now at the acceptable level.
-------------------------	--

REVIEWER	Cordia Chu Centre for Environment and Population Health, Griffith University, Australia
REVIEW RETURNED	15-May-2017

GENERAL COMMENTS	I have reviewed it already and the revision is acceptable.
-------------------------	--